4К

УДК 004.89:004.823

О.Б. Кунгурцев, С.М. Бородавкін
Одеський національний політехнічний університет, м. Одеса, Україна
sborodavkin@lonewolf.od.ua

Застосування мереж фреймів для побудови моделі вилучення фактів з текстів природною мовою

Розроблена модель вилучення фактів з текстів природною мовою та метод порівняння таких моделей з метою встановлення семантичної відповідності вхідних текстів. Даний метод може бути використаний при розробці інтерфейсу користувача інформаційної системи природною мовою, що дозволяє гнучко формувати запити до системи та синтезувати її вихідні повідомлення.

Вступ

При розробці інформаційних систем (ІС) основною задачею, що підлягає вирішенню, є задача пошуку інформації. Маючи порівняно просте рішення у випадку, коли інформація знаходиться у формалізованих структурах (РБД тощо), дана задача значно ускладнюється у випадку, коли ІС має порівнювати, аналізувати та синтезувати інформацію з порізнених фактів, розміщених у текстах природною мовою. Прикладами таких систем можуть служити пошукові системи, системи машинного перекладу, системи автоматизованої перевірки відповідей учня тощо. Іншою важливою задачею є задача класифікації об’єктів, якими оперує ІС. Чи відносяться до одного і того ж класу однаково пойменовані об’єкти різних БД? Чи про одне й те саме поняття йдеться в двох текстах природною мовою? При роботі з документами процес автоматичного структурування текстової інформації, поданої природною мовою, замінює експертний процес виділення фактів та об’єктів, що виконується вручну.

У даній роботі проводиться узагальнення підходу до вирішення даних задач та виконується формалізація математичної моделі їх вирішення. Розроблений метод аналізу текстів природною мовою та збереження знань (фактів, подій, об’єктів), поданих у ньому. Також розроблений метод порівняння моделей знань, отриманих із різних вхідних текстів.

Процес аналізу текстів природною мовою з метою виділення фактів (побудови моделі знань або семантичної моделі) поділяється на наступні етапи:

1. Синтаксичний розбір. На цьому етапі відбувається розбір речень вхідного тексту природною мовою з метою виділення їхніх членів та відношень між ними. Виконання цього етапу забезпечує: а) перевірку вхідного тексту на синтаксичну коректність, б) створення дерева розбору та підготовку даних для наступного етапу аналізу.

Задача синтаксичного розбору текстів природною мовою ускладнюється необхідністю підготовки словників (морфологічного словника, що містить граматичну інформацію про слова, словника оборотів, тощо), що, в свою чергу, потребує значних зусиль та є полем незалежних досліджень. Враховуючи це, а також наявність вже

готових рішень у даній сфері [1], [2], для виконання синтаксичного аналізу в даній роботі була вибрана система DictaScope, розроблена ТОВ «Dictum» [1]. Даний вибір обумовлений, в першу чергу, тим, що ця система дозволяє виконувати синтаксичний аналіз неповних мовних конструкцій (наприклад: «*Пішохід повинен керуватися сигналами пішохідного світлофору, а при його відсутності – транспортного світлофору*». Відновлене речення: «*Пішохід повинен керуватися сигналами пішохідного світлофору, а при відсутності [пішохідного світлофору] [пішохід повинен керуватися сигналами] транспортного світлофору*». У квадратних скобках наведені відновлені слова (заміна займенника «його» та відновлення еліптичної конструкції: у вхідному реченні тире замінює частину предикативної групи).

2. Семантичний аналіз. В рамках даного етапу на основі синтаксичного дерева розбору тексту природною мовою виконується побудова моделі знань, до якої подаються наступні вимоги:

1) повнота – модель повинна зберігати знання, закладені у неї, повністю, без викривлень та скорочень;

2) впорядкованість – модель повинна зберігати свою структуру при рості кількості «одиниць знань», закладених у неї, не перетворюючись на хаотичну множину елементів, пов’язаних між собою;

3) простота – елементи системи та зв’язки між ними повинні бути простими та доступними для аналізу людиною.

У ході аналізу різних когнітивних структур (а саме семантичних мереж і мереж фреймів), внутрішньою формою уявлення в системі була вибрана мережа фреймів [3], яка, на відміну від семантичної мережі, дозволяє більш упорядковано організувати базу знань системи і забезпечує впорядковування хаосу, властивого структурам на основі семантичної мережі. Фрейм одночасно містить великий обсяг знань і в той же час є достатньо гнучким для того, щоб бути використаним як окремий елемент бази даних [4]. Таким чином, як база знань може бути представлений текст природною мовою (набір текстів), що являє собою сукупність фреймоподібних одиниць (мережа логічно зв’язаних між собою фреймів і їх слотів, де зв’язками будуть певні відношення, що відображають взаємодію між об’єктами фреймів, їх ієрархію і характеристики).

Семантичний аналіз

Існує декілька підходів для побудови моделі знань на основі фреймів із вихідних даних синтаксичного розбору. Так, в роботі [5] пропонується використання граматики узагальнених складових (GPSG, Generalized Phrase Structure Grammar) з метою подання інформації з тексту природною мовою у вигляді формалізованих логічних форм, більш придатних для автоматизованого аналізу, показано також, що дана задача є NP-трудною; разом з цим існує підхід для перетворення GPSG-форм у мережу фреймів [6]. В роботі [7] описаний підхід до побудови фреймової мережі з тексту китайською мовою, полегшений попередньою розстановкою синтаксичних маркерів – ознак, що підкреслюють риторичні відношення між частинами речення (підпорядкування та координації). В цій роботі розстановка синтаксичних маркерів замінює синтаксичний розбір. Нарешті, в роботі [8] зображений підхід до побудови семантичної мережі з тексту англійською мовою та наведені рекомендації щодо її перетворення у мережу фреймів.

Однією з головних проблем семантичного аналізу, який виконується на основі попереднього синтаксичного розбору, є змістове поєднання речень між собою. Так, оскільки одне й те саме поняття (фрейм) може розкриватися (наповнюватися атрибу-

тами) протягом кількох речень, існує проблема того, як пов'язати, наприклад, іменник «*кіт*» із займенником «*він*» у парі речень «*Кіт сидить на вікні. Він дивиться на вулицю*». Навіть за відсутності безпосереднього синтаксичного зв'язку між конструкціями природної мови «*сидить на вікні*» та «*дивиться на вулицю*», в ході семантичного аналізу їх необхідно поєднати з фреймом «*кіт*». Такі зв'язки називаються анафоричними; те речення або його частина, в якому безпосередньо згадується об'єкт (назва фрейму), називається антецедентом; інше речення, в якому властивості (атрибути) фрейму розкриваються без прямого посилання на нього, називається анафором. Висловлювання, що включає анафор без антецеденту, навіть синтаксично завершене, має неповний зміст – у інших випадках зміст необхідно відтворювати через аналіз анафоричних зв'язків. Розробці автоматичного методу вирішення цієї проблеми присвячена робота [9]; крім того, система DictaScope, що використовується, має властивість часткового відновлення анафоричних зв'язків [1], а саме:

а) заміна іменника на прикметник («*Зелений сигнал світлофора дозволяє, а червоний [сигнал світлофора]– забороняє рух*»);

б) відсутність прямого додатку («*При ДТП водій зобов'язаний зупинити [транспортний засіб] та не рухати з місця транспортний засіб*»);

в) відсутність частини висловлювання при порівняльному прислівнику («*Машина має чотири колеса або більше [чотирьох коліс]*»);

г) заміна іменника на займенник;

ґ) еліптичні конструкції (заміна фрагмента складного речення на тире).

Модель вилучення фактів

У даній роботі фрейм розглядається як структура, що містить пойменовані елементи (слоти). Фрейми та їхні слоти наділяються певною семантикою, залежно від предметної області, в якій вони використовуються. Наприклад, фрейм може описувати деяке поняття, тоді слоти інтерпретуються як дії, властиві даному поняттю або пов'язані з ним. Задача вилучення фактів зводиться до:

1) виконання синтаксичного аналізу C вхідного тексту природною мовою, що може бути виражений перетворенням (1):

$$C: L \to T, \tag{1}$$

де L – вхідний текст природною мовою, T – дерево, або, у загальному випадку, ліс синтаксичного розбору (ліс виникає у випадку, коли синтаксичний аналіз може бути виконаний лише частково внаслідок неможливості відтворення анафоричних зв'язків, змістової неповноти або синтаксичної некоректності тексту);

2) виконання семантичного аналізу, результатом якого може виступати виявлення серед результатів синтаксичного аналізу T-присутності деякого фрейма F_i та визначення його слотів $\{S_{ij}\}$ відносно фрагментів дерева (лісу) синтаксичного розбору T. Крім того, для кожного слота S_{ij} (характеристики фрейма F_i) має значення кількість разів його появи у вхідному тексті N_{ij}. Той факт, що деякий слот S_{ij} зустрівся у фреймі N_{ij} разів, а інший слот S_{ik} – N_{ik} разів, дає змогу встановити глибину розкриття кожної з характеристик, вираженої відповідними слотами. Таким чином, частоту появи слота S_{ij} у фреймі F_i можна представити як

$$P_{ij} = \begin{cases} \dfrac{N_{ij}}{\left|\{S_{ij}\}\right|_i}, \left|\{S_{ij}\}\right|_i > 0 \\ 0, \left|\{S_{ij}\}\right|_i = 0 \end{cases}, \tag{2}$$

де N_{ij} – кількість разів появи слота S_{ij} у фреймі F_i, $\left|\{S_{ij}\}\right|_i$ – міцність множини слотів фрейма F_i.

Отже, семантичний аналіз полягає у формуванні для лісу синтаксичного розбору тексту, що аналізується, множини четвірок виду:

$$a_{ij} = \langle T_k, F_i, S_{ij}, N_{ij} \rangle, \quad (3)$$

де T_k – фрагмент дерева синтаксичного розбору, F_i – виявлений фрейм, S_{ij} – слот, що належить фрейму F_i та є виявленим значенням фрагмента T_i, N_{ij} – кількість разів появи слота S_{ij} у фреймі F_i. Отже, семантичний аналіз можна представити перетворенням виду:

$$Q: T \rightarrow \{a_{ij}\}. \quad (4)$$

Після того як множина четвірок a_{ij} сформована, можна здійснити перетворення виду:

$$P: a_{ij} \rightarrow a_{ij}^*,\ a_{ij}^* = \langle T_k, F_i, S_{ij}, P_{ij} \rangle, \quad (5)$$

замінивши абсолютні величини N_{ij} на відповідні їм відносні величини P_{ij}.

Можливий формальний опис структури фреймів та слотів наведений у роботі [4]. Виходячи з (1), (4), (5), модель вилучення фактів можна описати системою (6):

$$E = \langle C, Q, P \rangle. \quad (6)$$

Наведемо приклад виконання семантичного розбору тексту *«Попереду знаходиться відома вулиця. На ній мешкав Пушкін. Вулиця знаходиться у Одесі»*.

Множина трійок з (3), разом із частотою появи слотів (2), має вигляд:

$a_{11} = \langle$*”Попереду”, ПОПЕРЕДУ*, ∅, 0$\rangle$; $P_{11} = 0$
$a_{21} = \langle$*”знаходиться”, ЗНАХОДИТИСЬ*, S_{11}, 1$\rangle$; $P_{21} = 0.33$
$a_{22} = \langle$*”знаходиться”, ЗНАХОДИТИСЬ*, S_{41}, 2$\rangle$; $P_{22} = 0.66$
$a_{23} = \langle$*”у Одесі”, ОДЕСА*, ∅, 0$\rangle$; $P_{23} = 0$
$a_{31} = \langle$*”відома”, ВІДОМА*, ∅, 0$\rangle$; $P_{31} = 0$
$a_{41} = \langle$*”Вулиця”, ВУЛИЦЯ*, S_{31}, 1$\rangle$; $P_{41} = 1$
$a_{51} = \langle$*”мешкав”, МЕШКАТИ*, S_{61}, 1$\rangle$; $P_{51} = 1$
$a_{61} = \langle$*”Пушкін”, ПУШКІН*, ∅, 0$\rangle$; $P_{61} = 0$

Міцність даної множини дорівнює кількості слотів в усіх виявлених фреймах.

Порівняння фактів

При використанні моделі (6) постає проблема встановлення того, чи дві моделі E_1 *і* E_2 описують одну й ту саму множину фактів. Розв’язання даної проблеми дасть змогу вирішувати наступні задачі:

1. Задача пошуку інформації (модель E_1) у документах (модель E_2).
2. Задача класифікації (чи належить об’єкт, що описується моделлю E_1, до класу об’єктів, що описується моделлю E_2).
3. Задача верифікації синтезованої відповіді системи (природною мовою) відносно множини фактів, про яку система повідомляє користувача (наскільки правильно була синтезована відповідь системи природною мовою та наскільки повно вона відображає інформацію, що зберігається). Ця задача зводиться до задачі (1).

Задачу пошуку можна виконати шляхом накладання фреймових мереж моделей одна на одну та порівняння множин фреймів, їхніх слотів та відношень між ними. При цьому слід враховувати той факт, що викладення одних і тих самих фактів в двох різних вхідних текстах із використанням різних відмінків, типів речень, тощо, може призвести до появи відмінностей у фреймових мережах моделей. Щоб виключити дану проблему, в роботі [4] пропонується використовувати матрицю коефіцієнтів подібності семантичних відношень з метою нестрогого порівняння фреймових мереж із певною ймовірністю для кожного з відношень.

Задача 1 може бути зведена до пошуку значення функції відповідності між моделями E_1 *і* E_2:

$$R(E_1, E_2) = R_F(E_1, E_2) \times R_S(E_1, E_2)\,, \tag{7}$$

де R_F – функція відповідності між множинами фреймів обох моделей, R_S – функція відповідності між множинами слотів фреймів; обидві можуть приймати значення від 0 до 1. Таким чином, функція R також може приймати значення з [0;1], яке є характеристикою відповідності даних моделей.

Для визначення відповідності між множинами фреймів необхідно визначити ті з них, які є ключовими для порівняння (множини F_1^* і F_2^*). Це може зробити, наприклад, користувач або розробник системи. Враховуючи це, функція R_F може бути представлена у вигляді:

$$R_F(E_1, E_2) = r(F_1^*, F_2^*)\,, \tag{8}$$

де r – відношення між множинами ключових фреймів:

– якщо вирішується задача пошуку даних в моделі E_1 щодо запиту, поданого в моделі E_2, то r – відношення зворотнього включення ($F_1^* \supseteq F_2^*$);

– якщо виконується верифікація моделі E_1, тобто перевіряється, чи містить модель E_2 всі факти (фрейми) з E_1, то r – відношення прямого включення ($F_1^* \subseteq F_2^*$) тощо.

Визначимо функцію відповідності слотів R_S. Для цього відсортуємо слоти ($S_{i1}\ldots S_{in}$, n – кількість слотів фрейма) кожного фрейма в обох моделях у порядку зменшення їхньої частоти. Нехай $S_{i1}\ldots S_{ik}$, $k \le n$ – слоти, які є ключовими для визначення фрейма F_i, за відсутності яких не можна вважати даний фрейм визначеним. Частота $\overline{P}_{ik}$ останнього ключового фрейма буде виступати «порогом», тобто всі слоти із меншою частотою появи будуть при порівнянні проігноровані. Тоді функцію R_S можна представити у вигляді (9):

$$R_S(E_1, E_2) = \begin{cases} 1, \forall\, S_{ij}^1 \in S_1\, \exists\, S_{il}^2 \in S_2 : P_{ij}^1 \ge \overline{P}_{ik}^1\ \&\ P_{il}^2 \ge \overline{P}_{ik}^2 \\ 0, \text{в іншому випадку} \end{cases}, \tag{9}$$

де S_{ij}^1 – деякий ключовий слот із моделі E_1, S_{il}^2 – деякий ключовий слот із моделі E_2; $\overline{P}_{ik}^1$ – порогове значення частоти появи слотів у i-му фреймі з моделі E_1 ($\overline{P}_{ik}^2$, відповідно, – з моделі E_2); P_{ij}^1 та P_{il}^2 – частоти появи слотів S_{ij}^1 і S_{il}^2 відповідно. Функція приймає значення 1 тоді і тільки тоді, коли для кожного з ключових слотів в моделі E_1 знайдеться відповідний слот в моделі E_2, який також повинен бути ключовим; у всіх інших випадках функція приймає значення 0.

Вирішення задачі 2 класифікації (встановлення відповідності об’єктів, що описуються моделлю E_1, до класу об’єктів, що описується моделлю E_2) може бути виконане наступним чином. Нехай модель E_2 описує деякий клас об’єктів шляхом визна-

чення фрейма F та його слотів $\{Sj\}$. Подібно до (9), користувач чи розробник ІС визначають множину ключових слотів, які обов'язково повинні бути присутніми у фреймі з E_1 для того, щоб його можна було віднести до класу, визначеного в E_2. Відповідь на питання, чи належить фрейм F_i з моделі E_1 до класу, визначеного фреймом F, можна отримати шляхом порівняння множин їхніх ключових слотів. Для цього може бути використаний вираз (9).

Висновки

На основі аналізу структур, що використовуються для інтелектуальної обробки даних, була розроблена модель для вилучення і збереження фактів з текстів природною мовою. Використання даної моделі дозволить виконувати проектування інформаційних систем із природномовним інтерфейсом. Запропонований метод порівняння даних моделей між собою для вирішення задач пошуку і класифікації об'єктів.

Література

1. [Електронний ресурс]. – Режим доступу : http://www.dictum.ru. Матеріали з автоматичної обробки текстів.
2. [Електронний ресурс]. – Режим доступу : http://www.aot.ru. Матеріали з автоматичної обробки текстів.
3. Marvin Minsky. A Framework for Representing Knowledge / Minsky Marvin // The Psychology of Computer Vision. – 1975. – P. 211-277.
4. Пиринова Е.А. Средства представления и анализа ответов обучаемого на естественном языке / Е.А. Пиринова, А.Б. Кунгурцев, Н.А. Новикова // Нові інформаційні технології навчання в навчальних закладах України. – 2003. – № 9. – С. 236-242.
5. Ristad Eric Sven. Defining natural language grammars in GPSG / Eric Sven Ristad // Proceedings of the 24th annual meeting on Association for Computational Linguistics. – 1986. – P. 40-44.
6. Michael E. Cleary From logical forms to knowledge frames: an experiment on scientific text / E. Michael // Technical report NU-CCS-92-24. – College of Computer Science, Northeastern University, USA, 1992.
7. Ho H.C. Using Syntactic Markers and Semantic Frame Knowledge Representation in Automated Chinese Text Abstraction / H.C. Ho, B.K. T'sou, Y.W. Chan // Proceedings of PACFoCoL I (1993) : Pacific Asia Conference on Formal and Computational Linguistics. – 1993. – P. 122-131.
8. Lei Shi. An Algorithm for Open Text Semantic Parsing / Lei Shi, Rada Mihalcea // Proceedings of the ROMAND 2004 Workshop on «Robust Methods in Analysis of Natural Language Data». – 2004.
9. Ахренова Н.А. Нахождение анафорических связей при автоматическом анализе текста / Н.А. Ахренова. – Коломенский государственный педагогический институт, 2003.

А.Б. Кунгурцев, С.Н. Бородавкин

Применение сетей фреймов для построения модели извлечения фактов из текстов на естественном языке

Разработана модель извлечения фактов из текстов на естественном языке и метод сравнения таких моделей с целью установления семантического соответствия входных текстов. Данный метод может быть использован при разработке интерфейса пользователя информационной системы на естественном языке, что позволяет гибко формировать запросы к системе и синтезировать её выходные сообщения.

O.B. Kungurtsev, S.M. Borodavkin

Frame Networks Application to the Development of the Model of Facts Extraction from the Natural Language Texts

The model of the facts extraction from the texts on the natural language and the comparison method of such models are developed, with the purpose of establishing the semantic correspondence between the input texts. This method can be used for the development of the information system user interface based on the natural language, which allows a flexible creation of the queries to the system and synthesize its output messages.

Стаття надійшла до редакції 22.06.2009.